IOSR IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. II (Mar-Apr. 2014), PP 100-106
www.iosrjournals.org

Experimental study of Web Page Ranking Algorithms

Rachna Singh Bhullar, Dr. Praveen Dhyani
(Computer Science Department, Guru Nanak Dev University, Amritsar, Punjab, India)
(Banasthali University – Jaipur Campus, Jaipur, Rajasthan, India)

Abstract: *Today's life is very much influenced by the internet. Everything can be accessed by making a search on the Internet. Web browsers and web search engines are the two major components of web searching. Web browsers are the software applications and web search engines are the software itself. To achieve the goal of web searching web indexer plays a very important role of indexing the web pages by making use of link analysis algorithms. In this paper we are going to compare the three most commonly used link analysis algorithms which are page rank, weighted page rank and the hub & authority algorithms.*
Keywords: *Page Rank Algorithm, Weighted Page Rank Algorithm, Link Analysis Algorithm, Hub and Authority Algorithm.*

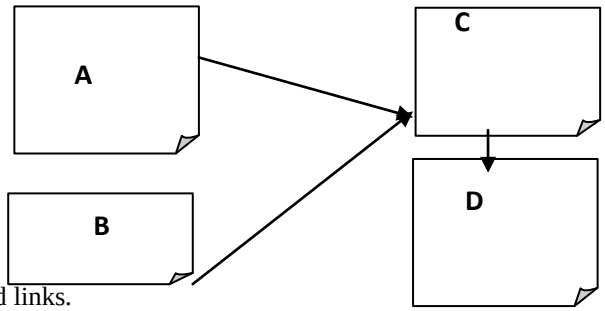
I. Introduction

In this computer era, Imagination of life without internet is impossible. To access the Internet, we need a web browser and a web search engine. Web Browser is application software which can be installed on the system to make the web search possible. Web Search Engine is software used by web browser to carry out the search on WWW. Examples of web browsers are Internet Explorer, Safari, Mozilla, Firefox, Opera, and Google Chrome whereas Doqpile, Webopedia, Google, Yahoo, AltaVista etc are the search engines. On the basis of the techniques used in searching, web search engines can be categorized as follows:

1. **Crawler Based Search Engines:** These types of search engines make use of spiders or bots to carry out their search. For Example, Google is crawler based search engine.
2. **Open Directory** such as Yahoo Directory, Open Directory and Look Smart are dependent on human editors to create their listings.
3. **Meta Search Engines** uses the results of other search engines to generate their searched list. For example, Doqpile and Mamma further serve the user supplied keywords to other individual search engines. Ultimately, they combine, integrate and sort the results.
4. **Hybrid Search Engines** Google and Yahoo both are Hybrid search engines because now a day's both are using the approach of crawler based and open directory based searching of web pages on WWW.

Hyperlink Structure of Web:

WWW is actually a massive directed graph in which nodes are the pages and directed links between the nodes are the hyperlinks. Inward links to a node are called Inlinks whereas Outward links are known as outlinks.



Types of links

a) Backlinks or Inlinks:
A and B are inlinks of C.
b) Outlinks or Forward Links:
C is A and B's outlinks.
D is C's Outlink.
c) **Dangling Links:**
1. Those links that point to any page having no forward links.
2. Those web pages which have not been downloaded yet.

Degree of webpage in WWW:

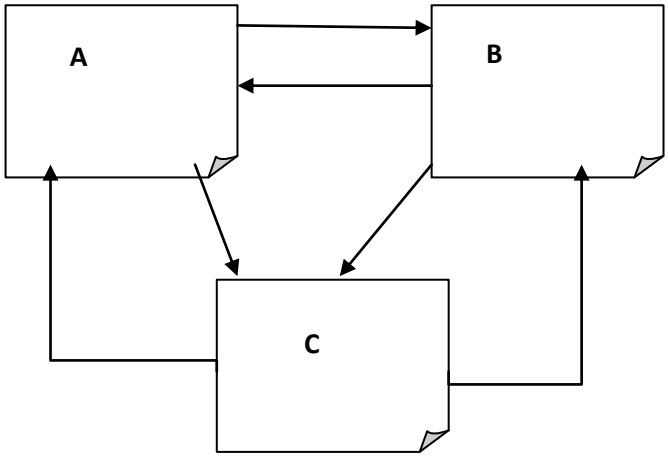
In an undirected graph, the number of links incident on a node is its degree.
But,
In directed graph, the degree of a node is of two types.

i. **Indegree:** It is the number of edges coming to the node is its Indegree.
For example,
Indegree(C)=2 and Indegree(D)=1

ii. Outdegree: It is defined as the number of edges going in outward direction from a node is its outdegree.
For example,
Outdegree(A)=Outdegree(B)=Outdegree(c)=1 and Outdegree(D)=0

Internal Linking:

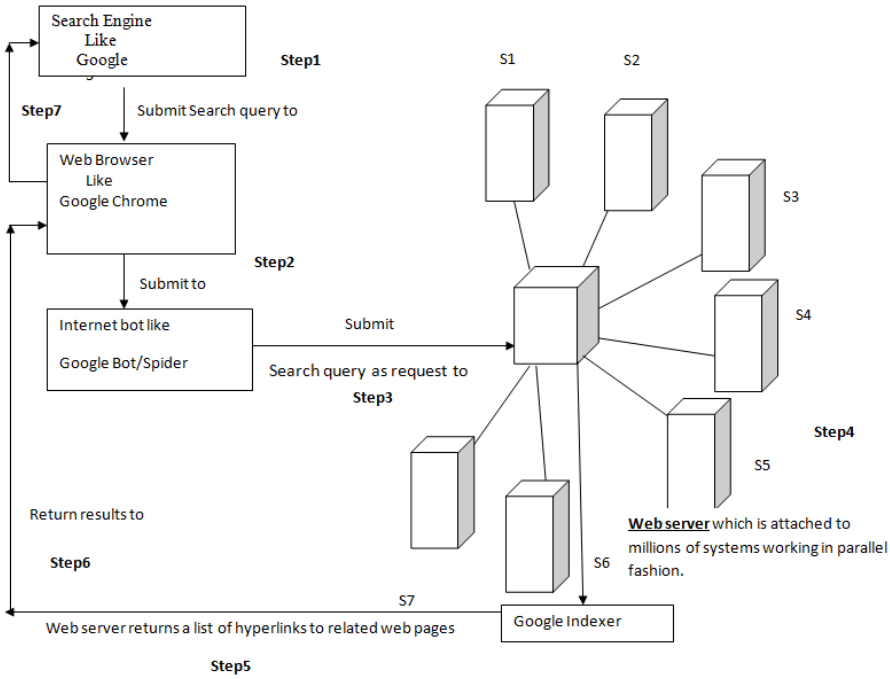
- Every page in the web has equal importance.
- Good Internal linking in a site would improve the page rank.
- Every page in the web graph is linked to other. i.e., web graph is a complete graph.



Outline of work
Section II provides the brief review of related work in web content mining. Section III explains the proposed work. Section IV provides the comparative study while in section V future work is summarized.

Section II

Related Work
This section mainly focuses on the working of the web search engine. Web Search Engine like Google makes use of Google bot or Internet bot to perform crawling. Indexing is performed by Google indexer and ultimately sorting is performed to obtain a list of pages in the result. The following diagram will explain the process of searching by web search engine.



The stepzwise description about the working of search engine is as follows:

Step1: Web Search Engine submits the query in separate keywords format to web browser.

Step2: Web browser gives the list of keywords to spider/ Internet bot like Google bot.

Step3: Google bot or internet spider crawls the whole network of WWW and get the links to the related pages. Actually, Internet Bot submit its query to web servers which in turn is connected to millions and millions of systems. Web Server then pass the keywords on the strands of network in WWW and get the list of related pages.

Step4: Web Server feeds the list of hyperlinks to related web pages to Google Indexer which then performs the indexing on the basis of various algorithms like Page Ranking, Weighted Page Rank, HITS algorithm etc.

Step5: The related web pages are ranked according to their usage and then sorted as well as indexed list is given to the Google browser.

Step6: Web browser converts the HTML format to user friendly format and submits the resultant web page to search engine.

Section III

Proposed Work: This section is going to describe the experimental evaluation of various link analysis algorithms as explained below.

Link Analysis Algorithms:

WWW is a huge graph of millions and millions of nodes linked to each other. Link analysis algorithms are one of the kinds of data analysis or data mining algorithms used to analysis the relationships between web pages and their importance in searching. We have Page Rank, Weighted Page rank and HITS as the various link analysis algorithms. These are explained as follows.

Page Rank Algorithm:

Page Rank was invented by Larry Page and Sergey Brin at Stanford University in their research paper which was later on adopted by Google to index their web pages.

1. Page Rank (PR) algorithm was introduced to rank the importance of web page in the hyperlink structure of web having millions of Inlinks and outlinks to other web pages.
2. PR actually is a numerical value that represents the importance of a web page.
3. When a web page, A refers to another web page (B), it casts a vote for that referred web page (B).
4. Vote (B) = $\sum A_{i,}$ referring to B.
   $i=1\ldots n$

Later on, Google implemented the algorithm in its indexing part of web searching.

Algorithm:

The PR algorithm invented by Larry Page & Sergey Brin is given by

$PR(A)=(1-d)+d(PR(T1)/C(T1)+\ldots\ldots\ldots\ldots\ldots\ldots+PR(T_n)/C(T_n))$

Where,

PR (A) =Page Rank of webpage, A.

PR (T_i) =Page Rank of pages T_i which link to Page (A).

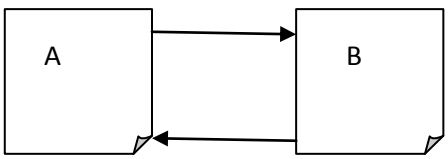
C (T_i) = number of outbound links on page T_i.

d =Damping factor which can have values between 0 & 1.

Or, in general,

Page Rank (PR) = 0.15+0.85*(a share of the page rank of every page that links to it) where d=0.85

Take an example of two web pages A & B to understand the concept of Page Rank:



1) If d=0.85
   PR (A) = (1-0.85) + 0.85* (PR (B) / C (B)), where, PR (B) =1 and C (B) =1
   PR (A) = 0.15+0.85=1
2) If d=0.85 and C(B)=1
   But, PR (B) =0
   PR (A) = 0.15+0.85*0=0.15

PR (B) =0.15+0.85*0.15 if C (A) =1
PR (B) =0.15+0.1275=0.2775
Let's say again perform the calculations as:
PR (A) = 0.15+0.85*0.2775=0.385875
PR (B) = 0.15+0.85*0.385875
= 0.5562946875
Again, PR (A) = 0.15+0.85*0.5562= 0.15+0.4675
= 0.6175
PR (B) = 0.15+0.85*0.6175
= 0.674875
Again, PR (A) = 0.15+0.85*0.674875
= 0.72364
PR (B) = 0.15+0.85*0.72364
= 0.7650

Again, PR (A) =0.80033
PR (B) =0.8302

Again, PR (A) =0.85574
PR (B) = 0.877379

Again, PR (A) =0.89577
PR (B) =0.91140

Again, PR (A) =0.9359
PR (B) =0.9455

Again, PR (A) =0.95375
PR (B) =0.96069

Again, PR (A) = 0.96658
PR (B) = 0.971598

Again, PR (A) = 0.97758
PR (B) = 0.97948

Again, PR (A) =0.98255
PR (B) =0.98517

Again, PR (A) =0.98739
PR (B) =0.98928

Again, PR (A) = 0.99089
PR (B) = 0.9922565

Again, PR (A) =0.993418

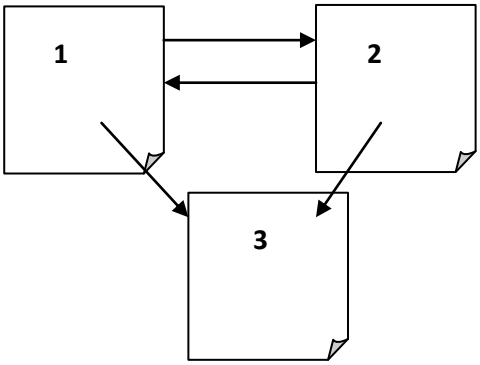
PR (B) =0.994405 a wrong value.

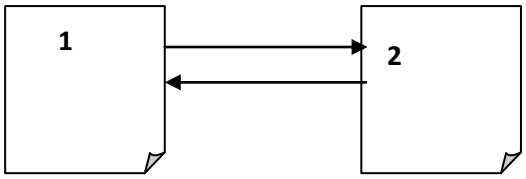
Again, PR (A) =2.3452, which is out of the range value.
Here, PR (A) =2.3452>1 which concludes that either our assumption of A ~~B~~ or PR (B) =0 Initial assumption is wrong.
This is a sample for the calculation of page rank of just two nodes. What if, we talk about the whole web of millions and millions of web pages and their page ranks, but something is there which seems to be wrong. Page and Sergey Brin, the founders of Google says that page rank is the average of all calculated Page ranks. Still, there are several questions to be answered by the implementation of this Page rank algorithm.

- What if, the initial assumption of any of T_i web page is wrong?
- What if, some web page is not referring to any other web page like $C(T_i)=0$, i.e., $PR(T_n)/C(T_n)=\infty$

- What if, damping factor d=0, PR (A) =1, whatever is the value of PR (T_n) and C (T_n).
- The above explained example of iteration for Page Rank shows the following two problems:

a) Rank sink b) Cycles





Weighted Page Rank:-

This algorithm was proposed by Weng U Xing and Ali Ghosbani as an extension of a Page Rank algorithm. Page Rank algorithm computes the rank on the basis of forward links of a webpage. Weighted Page Rank computes the rank by taking into all the importance of both back links and forward links of the webpage also distributes rank on the basis of popularity of the web page. In this algorithm more popular pages will have larger rank value than the value assigned by Page Rank algorithm. Hence in return to a user query, Weighted Page Rank return more relevant page than Page rank algorithm.

This algorithm computes rank as follows:

Step1:-

Computation of popularity from the number of inlinks as:

$$W^{in}_{(v,u)} = I_u / \sum_{p \in R(v)} I_p$$

Where,

W (v, u) = weight of link (v, u)

I_u =number of inlinks of page u

I_p =number of inlinks of page, p

R (v) =reference page list of page, v

Step2:-

Computes the weight of link (v,u) calculated on the basis of number of outlinks of page u and the number of outlinks of all reference pages of page, v as:

$$W^{out}(v, u) = O / \sum_{p \in R(v)} O_p$$

Where,

O_u and O_P represent the number of outlinks of page u and page p respectively. R(v) denotes the reference page list of page v.

Step3:-

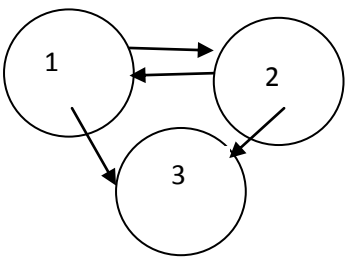
The original WPR formula is given as:

$$PR(u) = (1-d) + d \sum PR(v) W^{in}(v, u) W^{out}(v,u)$$

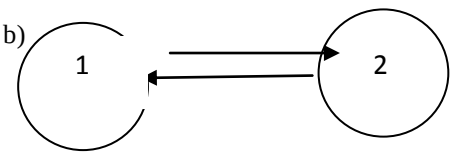
Implementation:-

Implement the WPR algorithm on Rank Sink and Cycle Problems of PR algorithm.

a)



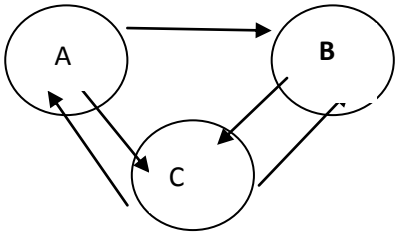
b)



Try to implement WPR algorithm on the case discussed in (b) part of the above figure.

1. If d=0.85 , W^{in} (B,A)=1/1 = No. of inlinks on A/ ∑(no. of inlinks on pages in reference list of page B)
   WPR (A) = (1-0.85) + 0.85*1= 0.15+0.85=1
2. Similarly, WPR(B)=1

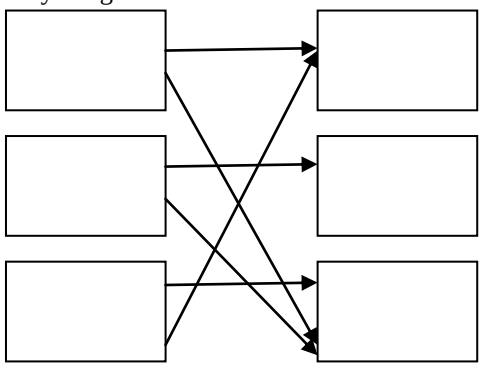
Thus, the unsolved problems of PR algorithm remain the same in WPR algorithm.
Let's consider one more example to compare the PR and WPR algorithm as shown in figure given below:



Ranks computed by PR algorithm are PR(C)>PR (B)>PR (A)
Whereas, ranks computed by WPR algorithm are WPR (A)>WPR(C)>WPR (B)

HITS (Hypertext Induced Topic Search) :-

HITS was proposed by Jon Kleinberg who was a young scientist at IBM in Silicon Valley and now a professor at Cornell University. Jon accepts a web page as a good hub if it points to worthy web pages and a good authority if it is pointed by the good hubs.



Jon started with a base set G, a set of pages in search resultant page, their incoming page links and outgoing page links. i.e. G=R U P U Q
Where, P is the set of related web pages
Q is set of forward links of P
R is set of forward links of P.
Then, algorithm is defined in the steps as:

Step1:-
Set each page authority as 1, initially.
Step2:-
Apply two update rules.

a) Authority Update Rule:
   ~~∀p~~ auth (p) = $\sum_{t=1\ldots n}$ hub (t)

   That is, authority score of a web page is equal to the summation of hub score of all its backlinks.

b) Hub Update Rule:
   ~~∀p~~ hub (p) = $\sum_{i=1..\ n}$ auth (i)
   That is, hub score of a web page is the summation of authority score of all its forward links.

Step3:-
Normalization

The above calculated authority and update scores are not the final scores as there is a need to converge them. Thus the concept of normalization is introduced to get the normalized authority and hub scores for each page.

Limitations of HITS algorithm:-

- A web page may be a hub as well as an authority, in that case to distinguish between hub pages and authority pages will be a difficult task.
- Some web pages having the same keywords and weights are actually not related to the user query, in that case HITS proves to be insufficient algorithm.
- HITS algorithm is not applicable in real time browsers.

Implementation-

HITS was implemented by IBM in a research project and used in proto-type search engine called Clever.

Section IV
Comparative Study of Link Analysis Algorithms

Algorithm	**Page Rank**	**Weighted Page Rank**	**HITS**
Mining technique used	WSM	WSM	WSM & WCM
Working	Computes the page importance at the time of indexing in Web Search Engine.	Computes the importance by assigning weight to its Back links and Forward Links.	Computes & Normalize the hub & authority scores.
I/P Parameters	Back links	In links & outlinks	In links, Out links& Content
Complexity	0(log N)	<0(log N)	<0 (log N)
Reliability	No solution for below problems :- (a)Rink Sink (b)Cyclic graphs.	No solution for below problems :- (a)Rink Sink (b)Cyclic graphs.	May solve :- (a)Rink Sink (b)Cyclic graphs.
Efficiency	Good when less no. of linked pages in web.	Better than Page rank algorithm in all cases.	Dependent of hub & authority behavior of page.
Web search engine	Goggle	Research Model	Clever

SectionV
Future Work

In the Section III part of the paper some questions are suggested at the end of the implementation phase of the Page Rank algorithm which are not answered by the implementation of WPR and HITS algorithm. This can be invented in future and one more question can be answered regarding the best of these three link analysis algorithms.

References:

[1]. file:///F:/thesis/Ian%20Rogers%20%C2%BB%20Google%20Page%20Rank%20%E2%80%93%20Whitepaper.htm
[2]. Study of Page Rank Algorithms, A Power Point Presentation.
[3]. Web Page Rank Algorithm based on number of visits of links of web page by Neelam Tyagi in International Journal of Soft Computing and Engineering(IJSCE), ISSN:2231-2307, Vol2, Issue3, July 2012.
[4]. A Survey Paper on Hyperlink Induced Topic Search Algorithm for Web Mining by Ramesh Prajapati in International Journal of Engineering Research and Technilogy, ISSN 2278-0181, volume 1, Issue 2, April2012,